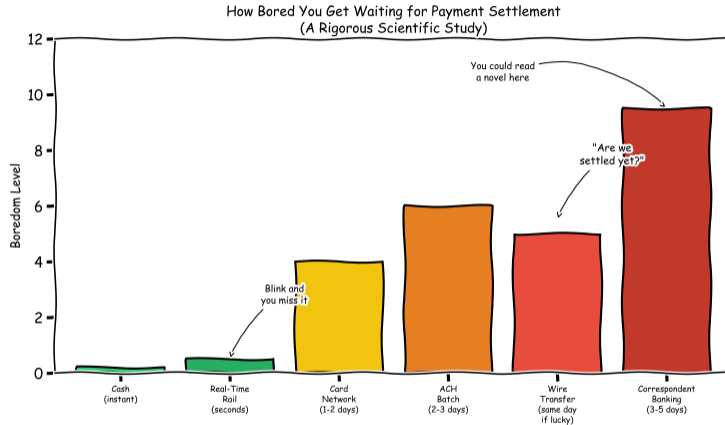


Lesson 1.3: Real-Time Payments and Cost Compression

Module 1: The Cost of Financial Intermediation

Digital Finance

The Long Wait for Settlement



Settlement speed varies dramatically across payment rails—from seconds to days.

By the end of this lesson, you will be able to:

- 1 Compare real-time payment architectures across different design patterns
- 2 Explain how ISO 20022 messaging standards enable interoperability
- 3 Analyze how stablecoin rails offer an alternative to traditional correspondent banking
- 4 Evaluate the economics of Buy Now Pay Later (BNPL) for merchants and consumers
- 5 Design a payment orchestration strategy that optimizes for cost, speed, and reliability

These objectives span Bloom's levels from comprehension through synthesis.

In Lessons 1.1–1.2, we learned why costs exist.

Information asymmetry → intermediaries → fees



Now we see how technology compresses them.

Real-time settlement, structured messaging, and programmable money can reduce layers, automate compliance, and shrink settlement windows from **days** to **seconds**.

Technology does not eliminate the need for trust—it changes who provides it and at what cost.

What is Real-Time Gross Settlement (RTGS)?

Definition:

Real-Time Gross Settlement (RTGS) is a funds-transfer system where each payment is settled *individually* and *immediately* on a gross basis, without netting or batching.

Key Properties:

- **Real-time:** Settlement occurs in seconds
- **Gross:** Each transaction settles individually (no netting)
- **Final:** Settlement is irrevocable once confirmed
- **Central bank money:** Settles on central bank ledger

Design Trade-offs:

- Requires participants to hold *liquidity buffers* at the central bank
- Higher per-transaction cost than batch systems
- Typically used for high-value / wholesale payments
- Examples: Fedwire (US), TARGET2 (EU), CHAPS (UK)

Contrast with Net Settlement:

- Batch systems net obligations at end of day
- Lower liquidity need, but *delayed* finality
- Introduces counterparty risk during the day

RTGS eliminates intraday credit risk but demands higher liquidity—a classic cost-risk trade-off.

Instant Payment Schemes: Architecture Patterns

What are Instant Payment Schemes?

Retail-focused systems enabling 24/7/365 real-time credit transfers between bank accounts, typically within seconds.

Design Pattern A: Centralized Clearing House

- Single operator validates and routes messages
- Participants pre-fund a settlement account
- Deferred net settlement at intervals (e.g., every 15 min)
- Example pattern: SEPA Credit Transfer Instant (SCT Inst)

Design Pattern B: Distributed / Hub-and-Spoke

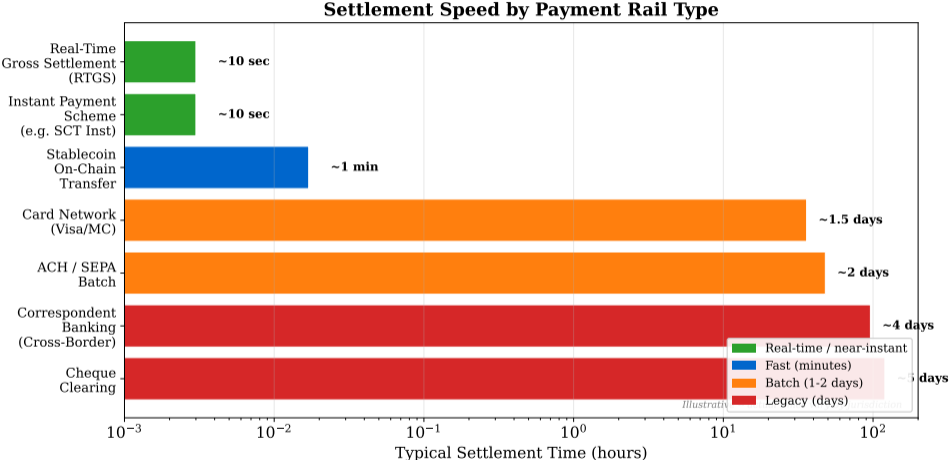
- Multiple interconnected clearing nodes
- Bilateral or multilateral settlement
- More resilient to single-point failure
- Example pattern: India's Unified Payments Interface (UPI design)

Design Pattern C: Overlay on Existing Rails

- Built atop legacy batch infrastructure
- Adds a real-time messaging layer
- Settlement still occurs on underlying RTGS
- Lower deployment cost, but legacy constraints remain

Different countries choose different architectures based on existing infrastructure and policy goals.

Settlement Speed: A Quantitative Comparison



Settlement speed spans five orders of magnitude—from seconds (RTGS) to days (correspondent banking).

What is Request-to-Pay (RtP)?

Definition:

Request-to-Pay (RtP) is a messaging framework that allows a payee to send a structured payment request to a payer, who can then approve, reject, or negotiate the payment.

How It Works:

- 1 Payee sends RtP message (amount, reference, due date)
- 2 Payer receives notification (mobile app, banking portal)
- 3 Payer reviews and *authorizes* payment
- 4 Payment executes over instant payment rail

Why It Matters for Cost Compression:

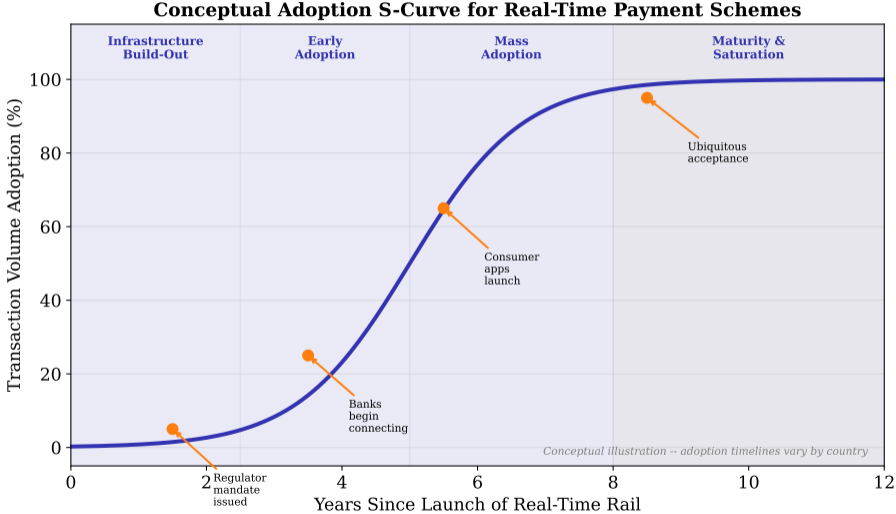
- Replaces direct debits (which have high dispute rates)
- Reduces failed-payment costs (payer confirms upfront)
- Enables flexible payment scheduling
- Rich data attached to each request (ISO 20022)

Use Cases:

- Bill presentment (utilities, subscriptions)
- E-commerce checkout (alternative to card payment)
- Business-to-business invoice settlement
- Government disbursements with confirmation

RtP shifts control to the payer while embedding rich data—reducing disputes and reconciliation costs.

Adoption Curves: How Real-Time Rails Scale



Real-time payment adoption follows an S-curve: slow build-out, then rapid growth once a critical mass of banks connect.

What is ISO 20022?

Definition:

ISO 20022 is a global standard for financial messaging that defines a common vocabulary and syntax (XML/JSON) for payment instructions, reporting, and securities settlement.

Key Features:

- **Structured data:** Named fields instead of free text
- **Rich remittance info:** Up to 140+ characters vs. 4 in legacy MT messages
- **Structured addresses:** Street, city, country as separate fields
- **End-to-end references:** Unique identifiers across the chain

Why It Enables Cost Compression:

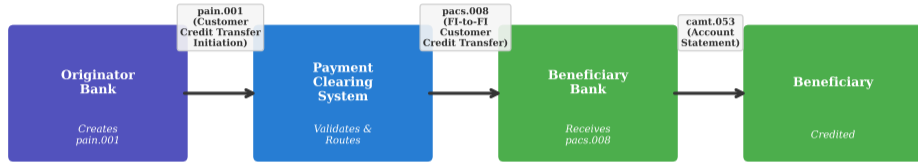
- **Straight-Through Processing (STP):** Machines can parse and route without human intervention
- **Fewer exceptions:** Structured data means fewer rejected or returned payments
- **Automated compliance:** Screening algorithms work on structured name/address fields
- **Interoperability:** Same message format across domestic and cross-border rails

Migration Timeline:

- SWIFT cross-border: migrating 2023–2025
- TARGET2 (EU): live since March 2023
- Fedwire (US): planned migration in progress

ISO 20022 is the “Rosetta Stone” of payments—enabling machines to process what humans used to reconcile manually.

ISO 2022: Structured Message Flow



ISO 2022 Data Richness (vs. Legacy MT/SWIFT Messages)

MT103: 4 remittance chars

-->

pain/pacs: 140+ remittance chars

Unstructured address fields

-->

Structured address (street, city, country)

Limited purpose codes

-->

Rich purpose codes + categories

No structured references

-->

Structured end-to-end references

From *pain.001* (initiation) to *pacs.008* (interbank transfer) to *camt.053* (reporting)—each message is structured and machine-readable

From Legacy MT to ISO 20022: What Changes?

Legacy MT Messages (SWIFT FIN):

- Fixed-length fields, positional syntax
- Free-text remittance (Field 70: 4 × 35 chars)
- Unstructured addresses
- Limited character set (ASCII subset)
- No end-to-end unique identifier

Result:

- High exception rate (~5–10% of cross-border payments)
- Manual repair of garbled addresses
- Compliance screening on unstructured text = false positives

ISO 20022 Messages:

- XML/JSON with named elements
- Structured remittance (140+ chars, typed fields)
- Structured addresses (street, building, city, country code)
- Unicode support
- Mandatory end-to-end transaction ID

Result:

- Near-zero exception rate for compliant messages
- Automated reconciliation (STP rates >99%)
- Compliance screening on structured fields = fewer false positives

The shift from MT to ISO 20022 is not cosmetic—it fundamentally changes what machines can do with payment data.

What are Stablecoin Payment Corridors?

Definition:

A stablecoin payment corridor uses blockchain-based tokens pegged to a fiat currency (e.g., USD, EUR) to transfer value across borders, bypassing traditional correspondent banking chains.

How It Works:

- 1 Sender deposits fiat with an on-ramp provider
- 2 Provider mints or transfers stablecoins on-chain
- 3 Stablecoins are sent to recipient's wallet (seconds)
- 4 Recipient uses an off-ramp to convert back to local fiat

Cost Compression Mechanisms:

- **Disintermediation:** Removes 2–4 correspondent banks
- **24/7 availability:** No batch windows or banking hours
- **Transparent fees:** On-chain gas fees are visible
- **Programmable:** Smart contracts can enforce compliance

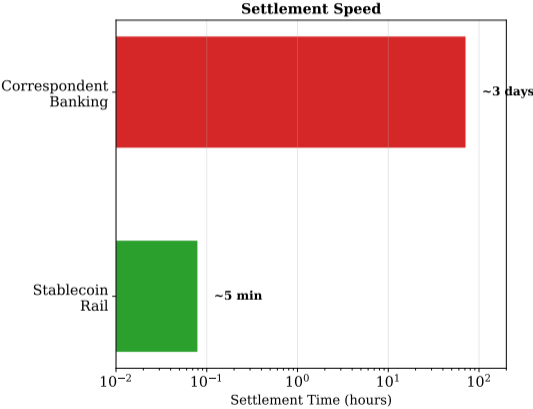
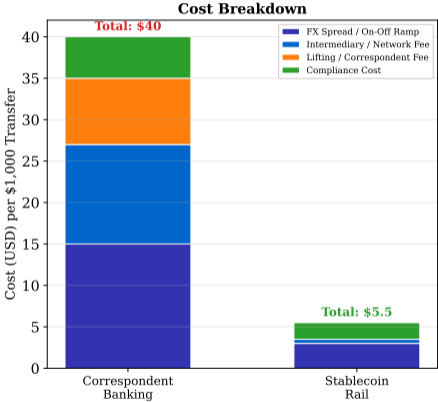
Challenges:

- On/off-ramp friction and regulatory licensing
- FX conversion still needed at endpoints
- Reserve transparency and audit requirements
- Regulatory uncertainty in many jurisdictions

Stablecoins compress the correspondent banking chain but introduce new risks at the on/off-ramp layer.

Stablecoin vs. Correspondent Banking: Cost and Speed

Cross-Border Payment: Correspondent Banking vs. Stablecoin Rails



Synthetic illustration – actual costs depend on corridor, volume, and provider

Stablecoin rails can reduce cross-border transfer costs and time dramatically—but on/off-ramp costs remain.

Stablecoin Corridor Design: Key Considerations

Chain Selection:

- Layer-1 (e.g., Ethereum): Higher security, higher gas cost
- Layer-2 (e.g., rollups): Lower cost, faster finality
- Purpose-built chains: Optimized for payments
- Trade-off: decentralization vs. throughput vs. cost

Reserve Model:

- Full fiat reserve (1:1 backing)
- Over-collateralized crypto reserve
- Algorithmic (no reserve—higher risk)
- Attestation vs. full audit transparency

Regulatory Requirements:

- Know Your Customer (KYC) at on/off-ramps
- Anti-Money Laundering (AML) transaction monitoring
- Travel Rule compliance (originator/beneficiary data)
- Licensing: e-money, payment institution, or banking

Corridor Economics:

- On-ramp fee: typically 0.1–1.0% of value
- Network/gas fee: **\$0.01–\$5** depending on chain
- Off-ramp fee: typically 0.1–1.5% of value
- Total: often <1% vs. 3–7% for legacy corridors

The “last mile” of stablecoin corridors—fiat on/off-ramps—is where most cost and friction remain.

What is Buy Now Pay Later (BNPL)?

Definition:

Buy Now Pay Later (BNPL) is a short-term consumer financing product that splits a purchase into installments (typically 3–4 payments), often with zero interest to the consumer. The merchant pays a Merchant Discount Rate (MDR) to the BNPL provider.

How It Works:

- 1 Consumer selects BNPL at checkout
- 2 BNPL provider runs a soft credit check (seconds)
- 3 Provider pays merchant immediately (minus MDR)
- 4 Consumer repays in installments over 4–6 weeks

Key Parties and Incentives:

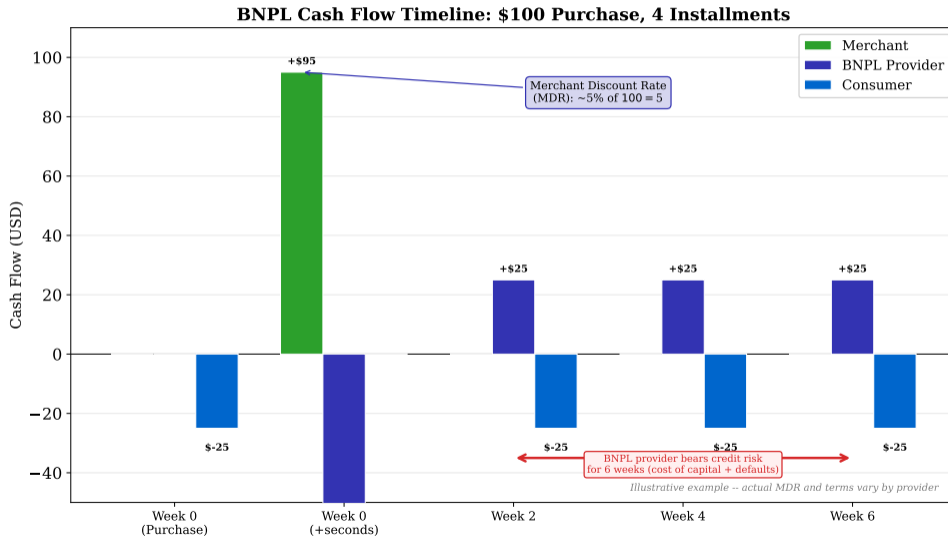
- **Consumer:** Interest-free installments → higher willingness to buy
- **Merchant:** Higher conversion rate, larger basket size → worth the MDR
- **BNPL Provider:** Earns MDR (~3–8% of transaction value)

Revenue Model for BNPL Provider:

- Primary: Merchant Discount Rate (MDR)
- Secondary: Late fees from consumers
- Tertiary: Interest on longer-term plans
- Data monetization (consumer spending patterns)

BNPL shifts credit cost from the consumer to the merchant—the merchant pays a higher MDR in exchange for increased sales.

BNPL Cash Flow Mechanics



BNPL Economics: Who Really Pays?

Merchant's Calculation:

- MDR for BNPL: typically 3–8% (vs. 1.5–3% for cards)
- Justified if conversion rate uplift \times basket size increase $>$ incremental MDR cost
- Example: If MDR increases from 2% to 5%, merchant needs a $>$ 3% revenue uplift to break even

BNPL Provider's Unit Economics:

- Revenue per transaction: MDR (e.g., 5% of \$100 = \$5)
- Cost of capital: funding the advance (\sim 0.5–1%)
- Credit losses: consumer defaults (\sim 1–4%)
- Operations: underwriting, servicing (\sim 0.5–1%)
- Contribution margin: thin (often $<$ 2%)

Consumer's Perspective:

- “Free” if paid on time—but is it?
- Late fees can exceed credit card interest
- Encourages spending beyond means
- No credit-card consumer protections (chargebacks)
- Multiple BNPL obligations may not appear on credit reports

Regulatory Response:

- EU Consumer Credit Directive: BNPL classified as credit
- UK FCA: bringing BNPL under regulation
- Affordability checks being mandated
- Disclosure requirements increasing

BNPL appears “free” to consumers, but merchants pay 2–3 \times card MDR—and default risk ultimately falls on the provider.

What is Payment Orchestration?

Definition:

Payment orchestration is a technology layer that sits between a merchant and multiple payment processors / acquirers, intelligently routing each transaction to optimize for cost, speed, and authorization success rate.

Core Functions:

- **Smart routing:** Select the lowest-cost or highest-success processor for each transaction
- **Failover:** If processor A declines, automatically retry with processor B
- **Multi-acquirer:** Connect to multiple acquirers in parallel
- **Payment method fan-out:** Cards, bank transfers, wallets, BNPL

Why It Matters:

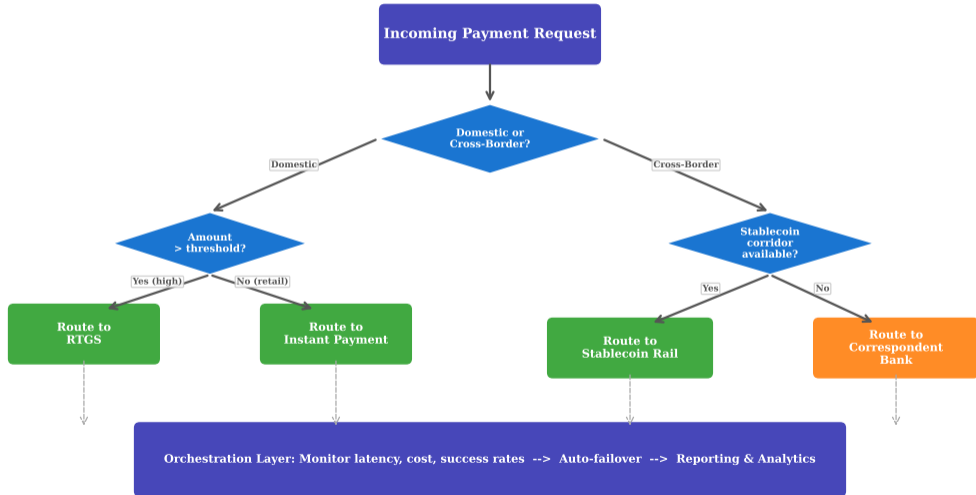
- A 1% improvement in authorization rate can yield millions in recovered revenue for large merchants
- Cross-border routing: local acquiring reduces FX markup
- Compliance: route to processors with correct licenses per jurisdiction

Orchestration Strategies:

- Cost-optimized: minimize processing fees
- Success-optimized: maximize authorization rate
- Speed-optimized: minimize settlement latency
- Balanced: weighted multi-objective routing

Payment orchestration turns payment processing from a single pipe into an intelligent, multi-path network.

Payment Orchestration: Intelligent Routing Decision Tree



Orchestration Metrics: What Gets Measured

Authorization Rate:

- Percentage of attempted payments that are approved
- Industry benchmark: 85–95% depending on sector
- Each 1% improvement \approx significant revenue recovery
- Affected by: BIN routing, 3DS version, retry logic

Cost Per Transaction:

- Interchange + scheme fee + acquirer markup
- Varies by card type, region, and MCC (Merchant Category Code)
- Orchestrator can route to minimize this
- Savings: 10–30 basis points per transaction

Settlement Latency:

- Time from authorization to funds received
- Ranges from T+0 (instant) to T+3 (batch)
- Cash-flow-sensitive merchants optimize for speed

Decline Analysis:

- Hard declines: insufficient funds, stolen card
- Soft declines: temporary holds, velocity limits
- Orchestrator retries soft declines on alternate rail
- Reduces false declines (legitimate transactions blocked)

Chargeback Rate:

- Disputed transactions as % of volume
- Threshold: typically $<1\%$ to avoid scheme penalties
- Routing can avoid high-chargeback corridors

Authorization rate is the single most impactful metric—a 1% lift can recover more revenue than a fee reduction.

What is Open Banking?

Definition:

Open banking is a regulatory and technological framework that requires banks to share customer account data and payment initiation capabilities with authorized third-party providers (TPPs) via standardized APIs, with the customer's explicit consent.

Regulatory Framework Types:

- **PSD2** (EU, 2018): Mandates API access; defines AISP and PISP roles
- **PSD3** (EU, proposed): Strengthens API performance standards, adds fraud-sharing
- **UK Open Banking**: CMA-mandated; standardized API specs
- **Market-led**: US, Australia—industry standards without mandate

Key Roles (PSD2/PSD3 Terminology):

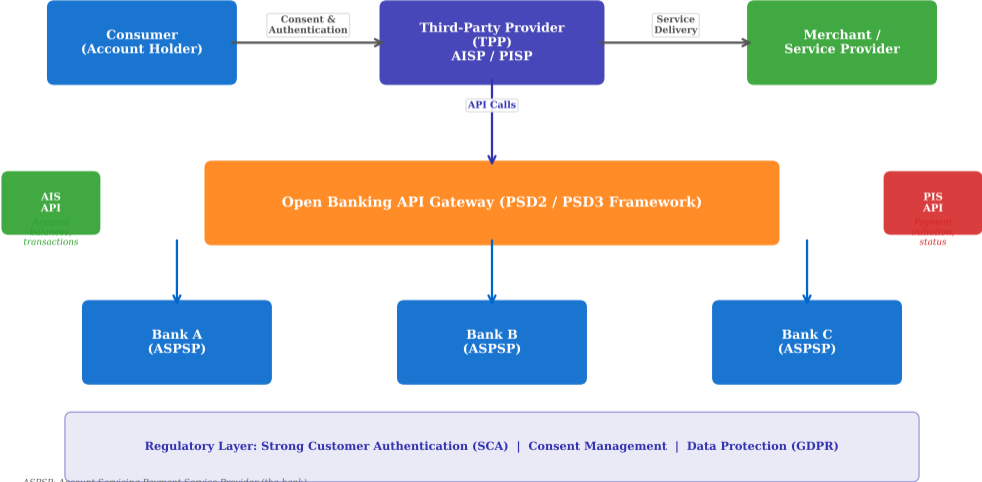
- **ASPSP**: Account Servicing Payment Service Provider—the bank
- **AISP**: Account Information Service Provider—reads balances/transactions (read-only)
- **PISP**: Payment Initiation Service Provider—initiates payments on behalf of the customer
- **TPP**: Third-Party Provider—umbrella term for AISP/PISP

Cost Compression Mechanism:

- Account-to-account (A2A) payments bypass card networks
- No interchange fee, no scheme fee
- Direct bank transfer at near-zero marginal cost

Open banking mandates API access to bank accounts—enabling account-to-account payments that bypass card networks entirely.

Open Banking: API-Based Data Flow Architecture



Open Banking: PSD2 vs. PSD3 Evolution

PSD2 (Second Payment Services Directive, 2018):

- Mandated banks to provide APIs to TPPs
- Introduced Strong Customer Authentication (SCA)
- Defined AISP and PISP licenses
- **Limitations encountered:**
 - API quality varied widely across banks
 - Screen-scraping fallback created security tension
 - No penalties for poor API performance
 - Fraud liability unclear for open-banking payments

PSD3 / PSR (proposed):

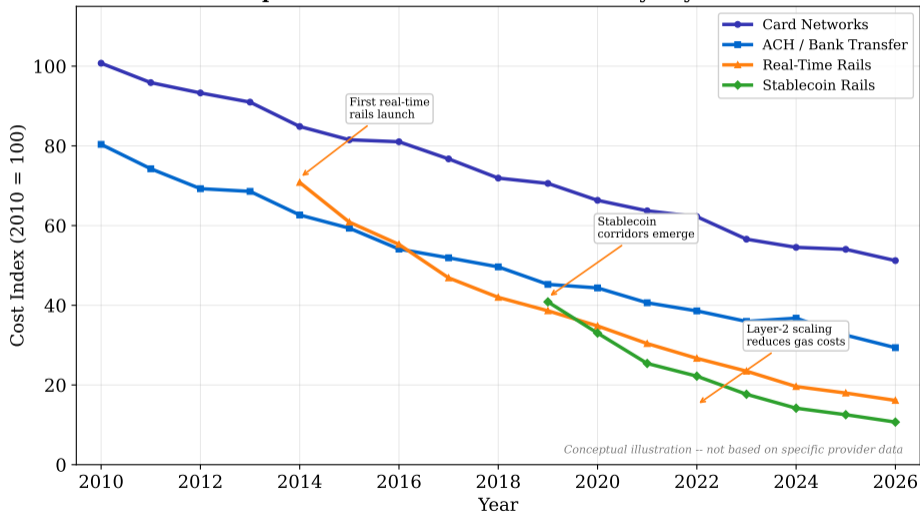
- Mandatory API performance standards with SLAs
- Dashboard for consumer consent management
- Fraud-data-sharing framework between banks and TPPs
- IBAN/name verification to prevent misdirected payments
- Direct access to payment systems for non-bank PSPs
- Removal of screen-scraping fallback

Net Effect on Costs:

- Higher-quality APIs → higher STP rates
- Fraud sharing → lower fraud losses
- Direct system access → lower intermediation cost

PSD3 addresses PSD2's implementation gaps—mandating API quality, fraud sharing, and direct system access for non-banks.

Conceptual Cost-Per-Transaction Trend by Payment Rail



Conceptual illustration – not based on specific provider data

The Coordination Problem: Why Adoption is Slow

Network Effects in Payments:

- A payment rail is only useful if both sender and receiver can use it
- Value scales with n^2 (Metcalfe's Law)—but so does the coordination cost
- Early adopters bear cost, late adopters get the benefit
- Result: chicken-and-egg problem

Incumbent Resistance:

- Banks earn interchange revenue from card networks
- Real-time A2A payments bypass this revenue stream
- Rational for incumbents to delay or under-invest in APIs
- Regulatory mandates (PSD2, UPI) overcome this

Strategies to Overcome Coordination Failure:

- **Regulatory mandate:** Government requires participation (EU: SCT Inst, India: UPI)
- **Public infrastructure:** Central bank builds and operates the rail
- **Killer use case:** One application drives mass adoption (e.g., QR-code merchant payments)
- **Subsidy:** Temporarily zero-fee to build volume, then monetize

Lessons from Adoption Patterns:

- Mandates accelerate adoption by 3–5 years
- Mobile-first design drives retail uptake
- B2B adoption lags retail by 2–3 years
- Interoperability across schemes remains unsolved

Payment adoption is a coordination game—technology alone is insufficient without regulatory push or a killer use case.

Comparing Cost Structures Across Rails

Payment Rail	Merchant Cost	Settlement	Availability	Data Richness
Card Network	1.5–3.5%	T+1 to T+3	24/7	Low (4-party model)
ACH / SEPA Batch	<0.5%	T+1 to T+2	Business days	Medium
Instant Payment Rail	<0.2%	Seconds	24/7/365	High (ISO 20022)
Open Banking (A2A)	<0.1%	Seconds–minutes	24/7/365	High
Stablecoin Corridor	0.3–1.5%	Seconds–minutes	24/7/365	Variable
BNPL	3–8% MDR	T+0 (merchant)	24/7	Medium

Key Observations:

- Instant rails and A2A are cheapest per transaction
- Card networks remain dominant due to consumer protection and ubiquity
- BNPL is the most expensive for merchants but increases conversion

Total Cost of Ownership:

- Per-transaction fee is only one component
- Integration cost, fraud liability, chargeback handling
- Reconciliation effort (high for legacy, low for ISO 20022)
- Cash-flow impact of settlement timing

The cheapest rail per transaction is not always the cheapest in total cost of ownership—consider integration, fraud, and reconciliation.

- 1 **RTGS and instant payment schemes** compress settlement from days to seconds, but require liquidity buffers and coordinated adoption
- 2 **ISO 20022** is the messaging standard that enables straight-through processing, reducing manual reconciliation and exception handling
- 3 **Stablecoin corridors** offer an alternative to correspondent banking, compressing cost and time—but on/off-ramp friction and regulation remain barriers
- 4 **BNPL shifts cost** from consumer to merchant via higher MDR; the provider bears credit risk and operates on thin margins
- 5 **Payment orchestration** intelligently routes transactions across multiple rails to optimize cost, speed, and success rate
- 6 **Open banking (PSD2/PSD3)** mandates API access, enabling account-to-account payments that bypass card-network fees entirely
- 7 **Adoption is a coordination problem:** technology alone is insufficient—regulatory mandates, public infrastructure, or killer use cases are needed to reach critical mass

Cost compression is technically possible today—the binding constraints are coordination, regulation, and incumbent incentives.

Real-Time Payments and Cost Compression

Core insight:	Technology <i>can</i> compress settlement from days to seconds
But:	Adoption faces coordination problems and incumbent resistance
RTGS / Instant Rails:	Real-time, gross, final settlement in central bank money
ISO 20022:	Structured messaging → straight-through processing
Stablecoins:	Bypass correspondent chains; on/off-ramp is the bottleneck
BNPL:	“Free” for consumers; merchants pay 3–8% MDR
Orchestration:	Intelligent routing across rails for cost/speed/reliability
Open Banking:	A2A payments bypass card networks at near-zero cost

Next: Lesson 1.4 — We examine how these building blocks combine in specific market segments and business models.

The gap between what is technically possible and what is widely adopted defines the opportunity space in digital payments.